Paper Presentation - Student Handout
Biostatistics 794, Fall 2025

Purpose. Develop the ability to *find, explain, and critically evaluate* a research article that applies **machine learning** (beyond standard linear/logistic regression) to a public-health problem, with attention to training/tuning/validation, interpretability, and ethics.

What papers are eligible?

- Must use an ML approach *or* a $p \gg n$ setting. Standard linear/logistic regression alone *does not* qualify.
- Examples of qualifying methods: decision trees, random forests, bagging/boosting (XGBoost/LightGBM), SVM, k-NN, naive Bayes, penalized regression (ridge/lasso/elastic net), neural networks/deep learning, ensemble stacking, clustering (k-means/hierarchical), dimension reduction (PCA/autoencoders), survival ML (e.g., random survival forests), text/NLP with embeddings, imaging CNNs.
- Prefer clear outcome definitions, transparent validation (CV or external test set), and strong public-health relevance (epi, health services, environmental health, genomics, neuroimaging, behavioral health).

Deliverables.

- **Proposal (due Sept 12):** Full citation + link; brief justification ($<$ 250 words) on why ML or $p \gg n$; data source/size; outcome/predictors; main method; how performance was estimated; any privacy/ethics notes.
- **Slides (PDF):** Upload 24 hours before presenting (presentations will start Sept 28th and run through Nov 18th).
- **Presentation:** ∼25 minutes in class (∼12–16 slides) + 5 minute discussion.

Suggested presentation structure (*guide*, not a script).

1. **Problem & Motivation:** why the question matters; *why ML was needed*: nonlinearity, interactions, complex data type, high dimensionality, prediction focus.
2. **Data & Design:** population; outcome; predictors; class balance; missingness; train/validation/test or external validation; leakage risks.
3. **Method Walkthrough:** conceptual explanation of the ML method; preprocessing/feature engineering; class imbalance handling; interpretability tools: variable importance, partial dependence plots.
4. **Training, Tuning & Validation:** what was tuned and how: grid/random/Bayesian search, early stopping; **CV or bootstrap**; metrics used and why: AUC/PR-AUC, calibration/Brier, RMSE/MAE.
5. **Results & Generalization:** main findings with uncertainty; subgroup/fairness checks; external/temporal validation; threats to validity.

6. **Ethics & Governance:** privacy/PHI, bias/fairness, transparency, reproducibility, code/data availability; fit to reporting guidance such as TRIPOD-ML, CONSORT-AI/SPIRIT-AI.
7. **Your Critique & Improvements:** what you trust/why; what you would change in data, design, method, validation, or deployment.

Logistics & tips.

- **Sign-ups:** class presentations will start Sept 28th and run through Nov 18th. I'll circulate a link to sign up for a date.
- **Slides:** Favor figures/diagrams over dense text; include 1 slide that clearly states *intended use, limits, and assumptions.*
- **R angle (encouraged):** Include a brief code sketch for a toy version of the method in `R` (e.g., `tidymodels`, `ranger`, `xgboost`, `glmnet`, `e1071`, `kernlab`, `iml/shapviz`).
- **Common pitfalls to address:** data leakage (features derived from outcome), tuning on the test set, reporting only accuracy for imbalanced data, ignoring calibration, no baseline comparator, no subgroup or site-level assessment, unclear outcome definition.

Good places to find papers. *Lancet Digital Health, Nature Medicine/Digital Medicine, npj Digital Medicine, JAMIA, AJPH, PLOS Digital Health, JMIR, Health Services Research (ML issues), Scientific Reports (health ML), Patterns, Bioinformatics.*

Questions? Bring candidate papers to office hours or post on the course forum for early feedback.

Reminder: *Your proposal is due Sept 12. Choose a paper that clearly uses ML (or $p \gg n$), reports validation carefully, and addresses a substantive public-health question.*

Table 1: Grading rubric (100 points)

Appropriateness of paper & justification	10
Problem framing & motivation for ML	10
Method explanation (conceptual clarity, correctness)	20
Tuning & validation strategy (methods, metrics, leakage awareness)	15
Results interpretation & baseline comparison	10
Interpretability, fairness, and ethics	10
Critical appraisal & generalization/transportability	10
Presentation quality (organization, visuals, timing, Q&A)	10
Reproducibility signals (reporting, code/data commentary)	5